

Simple Search Engine Model: Selective Properties

Mahyuddin K. M. Nasution^{1*}

¹Departemen Teknologi Informasi, FASILKOM-TI,
Universitas Sumatera Utara, Padang Bulan 20155 USU, Medan, Indonesia.
mahyunst@yahoo.com

Abstract. In this paper we study the relationship between query and search engine by exploring the selective properties based on a simple search engine. We used the set theory and utilized the words and terms for defining singleton and doubleton in the event spaces and then provided their implementation for proving the existence of the shadow of micro-cluster.

Keywords: singleton space, doubleton space, search term, query, triplet

1 Introduction

Numerous studies of natural language processing (NLP) and Semantic Web utilize a search engine, mainly to obtain a set of documents, mainly to obtain a set of documents that include a given query and to get statistical information about an object such as entity name in hit count [1] by the search engine, but to bring the NLP and Semantic Web to life such as the information processing services provide the knowledge, for example: ontology construction, knowledge extraction, question answering, and other purposes, all need more effort. However, to produce the enhanced relationship between a search engine and a query as novel property, we already defined some instances about simple search engines: singleton [2] and doubleton [3] spaces. This model based on the simple architecture of search engine for representing the collection of documents in general such that this model can distinguish the features of Web documents, whereby any query can give the essential purpose of Information Retrieval [4] strategies to meet user needs. Therefore, this paper aims to address some properties based on relations between singleton and doubleton in a triplet. We also provided the basis of this model to the micro-cluster for implementing the adaptive properties.

2 Some Terminologies

We defined some terminologies as follows [1,2,3].

* A draft

Definition 1. A term t_x consists of at least one or a set of words in a pattern, or $t_k = (w_1 w_2 \dots w_l)$, $l \leq k$, k is a number of parameters representing word w , l is the number of tokens (vocabularies) in t_k , $|t_k| = k$ is size of t_k . ■

Definition 2. Let a set of web pages indexed by search engine be Ω , i.e., a set contains ordered pair of the terms t_{k_i} and the web pages ω_{k_j} , or (t_{k_i}, ω_{k_j}) , $i = 1, \dots, I$, $j = 1, \dots, J$. The relation table that consists of two columns t_k and ω_k is a representation of (t_{k_i}, ω_{k_j}) where $\Omega_k = \{(t_k, \omega_k)_{ij}\} \subset \Omega$ or $\Omega_k = \{\omega_{k_1}, \dots, \omega_{k_j}\}$. The cardinality of Ω is denoted by $|\Omega|$. ■

Definition 3. Let t_x is a search term, and $t_x \in \mathcal{S}$ where \mathcal{S} is a set of singleton search term of search engine. A vector space $\Omega_x \subseteq \Omega$ is a singleton search engine event (singleton space of event) of web pages that contain an occurrence of $t_x \in \omega_x$. The cardinality of Ω_x is denoted by $|\Omega_x|$. ■

Definition 4. Let t_x and t_y are two different search term, $t_x \neq t_y$, $t_x, t_y \in \mathcal{S}$, where \mathcal{S} is a set of singleton search term of search engine. A doubleton search term is $\mathcal{D} = \{\{t_x, t_y\} : t_x, t_y \in \Sigma\}$ and its vector space denoted by $\Omega_x \cap \Omega_y$ is a double search engine event (doubleton space of event) of web pages that contain a co-occurrence of t_x and t_y such that $t_x, t_y \in \omega_x$ and $t_x, t_y \in \omega_y$, where $\Omega_x, \Omega_y, \Omega_x \cap \Omega_y \subseteq \Omega$. ■

Some adaptive properties are defined to know the efficient ways to access information by using simple search engine model. In general, all adaptive properties is to adopt the meaning of singleton and doubleton in equations as follows

$$|\Omega_x| = |\Omega_x| + |\Omega_y|$$

and

$$|\Omega_x \cap \Omega_y| = |\Omega_x \cap \Omega_y| + |\Omega_x \cap \Omega_x| + |\Omega_y \cap \Omega_y|$$

Therefore, statistically either singleton or the doubleton contain bias information.

Usually to improve the quality of statistical information by a search engine of a given query, the count is processed statistically based on above properties. However, to make an additional improvement, we must devote more attention to results of search engine and carefully handle the count for developing the selective model.

3 The Selective Properties

The purpose of selective properties is to construct an approach for eliminating bias by using the selected results of simple search engine. One of results by a search engine as follows [5].

Definition 5. Let t_x is a search term. $S = \{w_1, \dots, w_{max}\}$ is a Web snippet (briefly snippet), $S \subset \omega_{x_i} \in \Omega$, where $max \leq 50$ words to the left and right of t_x that returned by any search engine. $L = \{S_i : i = 1, \dots, n\}$ is a list of snippets. ■

3.1 Triplet

A construction of relationship based on frequency of words between search term, snippets, and words is as follows [6].

Definition 6. A relationship between search term, web snippets and words is defined as the mixture $p(t_o, S, w) = t_o \times S \times w$, $t_o \in O$, $S \in L \subseteq \Omega$, $w \in S$. A vector space of $P(t_o, S, w)$ is defined as $\mathbf{w} = \{w_i, \dots, w_j\}$. ■

$= [\nu_i, \dots, \nu_j]$, $\nu_i \geq \dots \geq \nu_j$, where w_i, \dots, w_j are the unique words in S and ν_i, \dots, ν_j are the weights of word.

The relations of the search term and the Web snippets and the words, we called it as triplet, or we rewrote as a term-snippet-word. The triplet is a base for exploring features of: Web pages or Web documents. The features exploration is to describe an object literally in text if the purpose of search term is to explain the object. A relation between term and snippet logically is $|t_o \cap S| = 1$ if $t_o \in S$ and $= 0$ otherwise, and

$$P(t_o \cap S) = \frac{1}{2}, \quad (1)$$

or probability of the search term in list of snippets are

$$P(t_o \cap L) = \frac{\sum_{i=1}^n \frac{1}{2}}{n}. \quad (2)$$

A relation of snippet-word interpreted as follows,

$$P(S \cap \mathbf{w}) = \sum_{j=1}^m \frac{1}{max} \quad (3)$$

where m is a number of same word in vocabulary, or probability of the word in list of snippets is as follows

$$P(L \cap \mathbf{w}) = \sum_{i=1}^n \sum_{j=1}^m \frac{1}{max_i}, \quad (4)$$

while the term-word has two representations logically, i.e. $|t_o \cap w| = \sum_{j=1}^m \frac{1}{max}$ if $t_o \in S$ and $= 0$ if $t_o \notin S$, or probability of $t_o \cap w$ in S as follows

$$P(t_o \cap \mathbf{w}) = \frac{\sum_{j=1}^m \frac{1}{max}}{2}. \quad (5)$$

Probability of $t_o \cap w$ in L to be

$$\begin{aligned} P(t_o \cap \mathbf{w})_L &= \sum_{i=1}^n \frac{\sum_{j=1}^m \frac{1}{max}}{2} \\ P(t_o \cap \mathbf{w})_L &= \sum_{i=1}^n \sum_{j=1}^m \frac{1}{2 max_i} \end{aligned} \quad (6)$$

Trivially for each snippet there exists a set of words $\mathbf{w} = \{w_i | i = 1, \dots, n\}$, i.e. \mathbf{w} contains at least one word of search term or the search term self.

Definition 7. Word frequency is a word number uniquely in a set of words, i.e. $\exists \nu \in \mathbb{R} \forall w \in \mathbf{w}$, \mathbb{R} is a real number set, and $\nu \in \mathbb{R}$ as a weight of word. Generally, there is 1 : 1 function ϖ such that

$$\varpi : \mathbf{w} \rightarrow \mathbb{R} \quad (7)$$

In this case, \mathbb{R} as a vector space of \mathbf{w} . ■

Lemma 1. If a set of words is representation of snippets in list of snippets, then vector space of words set contains probability of word in snippets.

Proof. In Eq. (3) as probability of word based on frequency of word in snippet where m is number of word uniquely in snippet. Therefore, Eq. (4) also is probability of word based on frequency of word in list of snippets. Reasonably, because $|t_o \cap \mathbf{w}| \neq 0$. Based on Eq. (6) and Eq. (7) we have

$$\nu = \varpi(w) = \sum_{i=1}^n \sum_{j=1}^m \frac{1}{2 \max_i}$$

as probability of word in a list of snippets for a search term, and \mathbb{R} contains the value of Eq. (6) for all $w \in \mathbf{w}$. ■

Lemma 2. If a set of words is representation of snippets in list of snippets, then the set of words is an event space.

Proof. As a search term each word in \mathbf{w} based on Definition 3 has Ω_w , and each two words in pair based on Definition 4 has $\Omega_{w_i} \cap \Omega_{w_j}$. Thus \mathbf{w} be an event space. ■

The event space contains vectors $\mu_i = |\Omega_{w_i}|$, $i = 1, \dots, n$, and we called it as singleton event space.

Proposition 1. If $p(t_o, S, w)$ is a triplet for t_o , then there are at least one vector space of $p(t_o, S, w)$.

Proof. The direct consequence of Lemma 1 and Lemma 2. ■

Thus, based on Lemma 1 and Lemma 2, there are two vector spaces for a list of snippets, i.e.

Definition 8. A set of words \mathbf{w} is a context if \mathbf{w} has two vector spaces such that satisfies

1. for $[\nu_i, \dots, \nu_j]$ as vector space, $\nu_i \geq \dots \geq \nu_j$, and
2. for $[\mu_i, \dots, \mu_j]$ as singleton vector space, $\mu_i \geq \dots \geq \mu_j$. ■

3.2 Micro-cluster

In part of section, we define the words undirected graph $G = (V, E)$ to describe the relations between words [7].

Definition 9. Assume a sub-graph G' , $G' \subset G$, G' is a micro-cluster satisfies the conditions as follows

1. There are a set of word $\mathbf{w} = \{w_i, \dots, w_j\}$ whose vector space $[\nu_i, \dots, \nu_j]$ and $\nu_i \geq \dots \geq \nu_j \geq \alpha$, where α is a threshold.
2. There are an one-one function $f : \mathbf{w} \rightarrow V$ such that $f(w) = v$, $\forall w \in \mathbf{w} \exists v \in V$ where $v \in V$ is a vertex in G' .
3. There are an one-one function $\rho : \mathbf{w} \times \mathbf{w} \rightarrow E$ such that $\rho(w_i, w_j) = e$, $\forall w_i, w_j \in \mathbf{w}$, where ρ is a relation among words and $e \in E$ is a edge in G' .

The micro-cluster is denoted by $G' = \langle V, E, \mathbf{w}, f, \rho, \alpha \rangle$. ■

A micro-cluster is maximal clique sub-graph of entity name where the node represents word that the highest score in document. However, let there is a set of words \mathbf{w} whose weights above the threshold, the collection of words do not exactly refer to the same entity. To group the words into the appropriated cluster, we construct the trees of words. This based on an assumption that the words are that appear in same domains having closest relation. The tree is an optimal representation of relation in graph G .

Definition 10. A tree T is an optimal micro-cluster if and only if T is a sub-graph of micro cluster G' , and is denoted by $T = \langle V_T, E_T, \mathbf{w}_T, f, \rho, \alpha \rangle$, where $V_T \subseteq V$, $E_T \subseteq E$, and $\mathbf{w}_T \subseteq \mathbf{w}$. ■

In building the optimal micro-cluster, we save the strongest relations in T between a word and another in G' until T has no cycle. A cycle is a sequence of two or more edges $(v_i, v_j), (v_j, v_k), \dots, (v_{k+1}, v_i) \in E$ such that there is an optimal edge $(v_i, v_j) \in E$ connects both ends of sequence. Let a word is introduced as intrusive word about an entity, and there are at least one word of optimal micro-cluster has strongest relations with the entity, and an optimal micro-cluster is a group of words refer to that entity. However, the overlap keyword also exists in the same list. We define a strategy to select a relevant keywords among all list candidates. In this case, there are a few potential keywords for identifying the entity name.

Definition 11. A vector space $\mathbf{s} = [|\mathbf{w}_i|, \dots, |\mathbf{w}_j|]$ is a mirror shade of micro-cluster G' if there is an one-one function $g : \mathbf{w} \rightarrow \mathbf{s}$, where $\mathbf{w}_i, \dots, \mathbf{w}_j$ are in event space. Let \mathbf{z} is a vector whose greatest value in \mathbf{s} , the vector space in range of $[0, 1]$ is relatively defined as $\mathbf{s}_{[0,1]} = [|\mathbf{w}_i|/|\mathbf{z}|, \dots, |\mathbf{w}_j|/|\mathbf{z}|] = [\mu_i, \dots, \mu_j]$. ■

We also can generate for example another vectors from $\Omega_i, \dots, \Omega_j$ for words w_i, \dots, w_j respectively such that $[\mu_i, \dots, \mu_j] = [\Omega_i, \dots, \Omega_j]$ is a mirror shade of $[\nu_i, \dots, \nu_j]$ from a set of word frequencies.

Theorem 1. *Let $s_T \subseteq s$, then s_T is the mirror shade of an optimal micro-cluster T .*

Proof. Let $s_T \subseteq s$, based on Definition 11 we have $w_T \subseteq w$, i.e. $g(w_T) = s_T$ or because of g is one-one function, $g^{-1}(s_T) = w_T \subseteq w$. Next, by applying Definition 9, $f(w_T) = V_T$, or because of f is one-one function, $f^{-1}(V_T) = w_T \subseteq w$, and $s_T = g(w_T) = g(f^{-1}(V_T)) = f^{-1}g(V_T)$ and we obtain $\rho(w, w) = \rho(f^{-1}(V), f^{-1}(V)) \subseteq E$, so $\rho(s_T \times s_T) = \rho(g(w_T) \times g(w_T)) = \rho(g(f^{-1}(V_T)) \times g(f^{-1}(V_T))) = \rho(f^{-1}g(V_T) \times f^{-1}g(V_T)) = f^{-1}g\rho(V_T \times V_T) = f^{-1}g(\rho(V_T \times v_T))$ because of $f^{-1}g$ is also one-one function, this means that $V_T \subseteq V$ has s_T as a mirror shade of w_T . ■

4 Conclusions and Future Work

The selective properties have been derived from the singleton and doubleton based on the triplet concept. Through these properties have been proven the existence of the shadow of any micro-clusters for the space of events. Our future work is about the relation between adaptive and selective properties for exploring an overlap principle.

References

1. M. K. M. Nasution. Kolmogorov complexity: Clustering and similarity. *Bulletin of Mathematics*, 3(1): 1-16, 2011.
2. M. K. M. Nasution. Simple search engine model: Adaptive properties. arXiv:1212.3906, Cornell University Library: 2012.
3. M. K. M. Nasution. Simple search engine model: Adaptive properties for doubleton. arXiv:1212.4702v1, Cornell University Library: 2012.
4. M. K. M. Nasution and S. A. Noah. Information retrieval model: A social network extraction perspective. In *IEEE Proc. of CAMP 2012*: 322-326, 2012.
5. M. K. M. Nasution, S. A. Noah, and S. Saad. Social network extraction: Superficial method and information retrieval. In *Proceeding of International Conference on Informatics for Development (ICID'11)*: c2-110-c2-115, 2011.
6. M. K. M. Nasution and S. A. Noah. Probabilistic generative model of social network based on web features. *Proceedings of International Seminar on Operation Research (InteriOR 2011)*: 241-250, 2011. Or arXiv:1207.3894v1, Cornell University Library: 2012.
7. M. K. M. Nasution and Shahrul Azman Noah. Superficial method for extracting social network for academics using web snippets. *Rough Set and Knowledge Technology*, LNCS - LNAI 6401, Springer-Verlag: 483-490, 2010.